DATA ANALYSIS I

Data: Algebraic and Geometric View

2019-20

Sources

- Zaki, M. J., Meira Jr, W. (2014). Data Mining and Analysis: Fundamental Concepts and Algorithms. Cambridge University Press. [1-13]

Data Matrix

- Data can be represented or abstracted as an $n \times d$ data matrix, with n rows and d columns.
- Rows correspond to objects in the dataset.
- Columns represent attributes or properties of interest.

Data matrix

$$
\mathbf{D} = \left(\begin{array}{c|cccc}
 & X_1 & X_2 & \cdots & X_d \\
\hline
\mathbf{x}_1 & x_{11} & x_{12} & \cdots & x_{1d} \\
\mathbf{x}_2 & x_{21} & x_{22} & \cdots & x_{2d} \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
\mathbf{x}_n & x_{n1} & x_{n2} & \cdots & x_{nd}
\end{array}\right)
$$

$$\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{id})$$

$$X_j = (x_{1j}, x_{2j}, \ldots, x_{nj})$$

Data Point

- If the *d* attributes or dimensions in the data matrix *D* are all numeric, then each row can be considered as a *d*-dimensional point

$$\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{id}) \in \mathbb{R}^d$$

$$\mathbf{x}_i = \begin{pmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{id} \end{pmatrix} = \begin{pmatrix} x_{i1} & x_{i2} & \cdots & x_{id} \end{pmatrix}^{\mathrm{T}} \in \mathbb{R}^d$$

Cartesian Coordinate Space

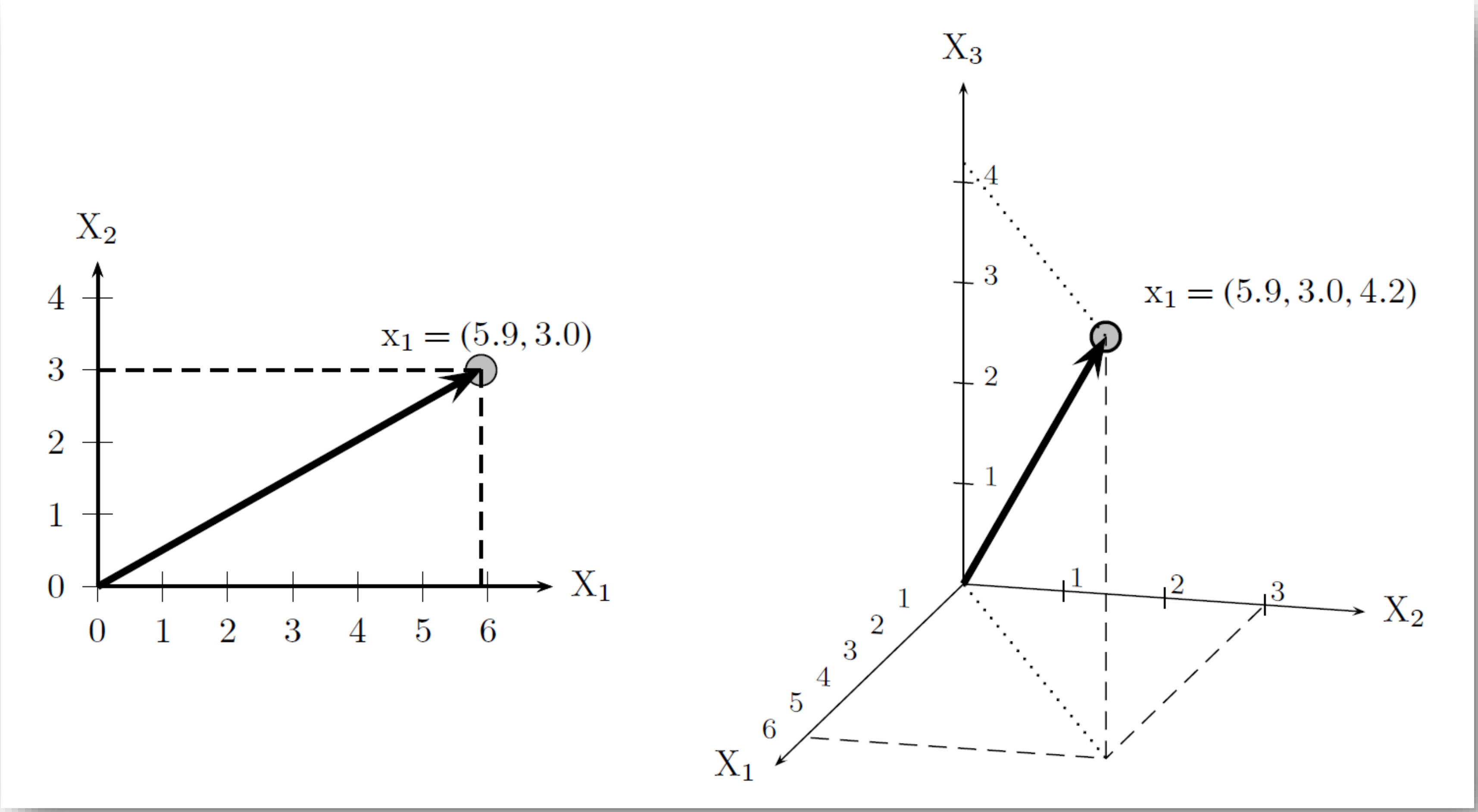
- The *d*-dimensional *Cartesian coordinate space* is specified via the d unit vectors, called the standard basis vectors, along each of the axes.

- Any other vector in R^d can be written as linear combination of the standard basis vectors.

$$\mathrm{e}_j = (0, \ldots, 1_j, \ldots, 0)^{\mathrm{T}}$$

$$\mathrm{x}_i = x_{i1}\mathrm{e}_1 + x_{i2}\mathrm{e}_2 + \cdots + x_{id}\mathrm{e}_d = \sum_{j=1}^{d} x_{ij}\mathrm{e}_j$$

Examples



Data Matrix: Example

- An extract of the Iris dataset; the complete data forms a *150×5* data matrix.

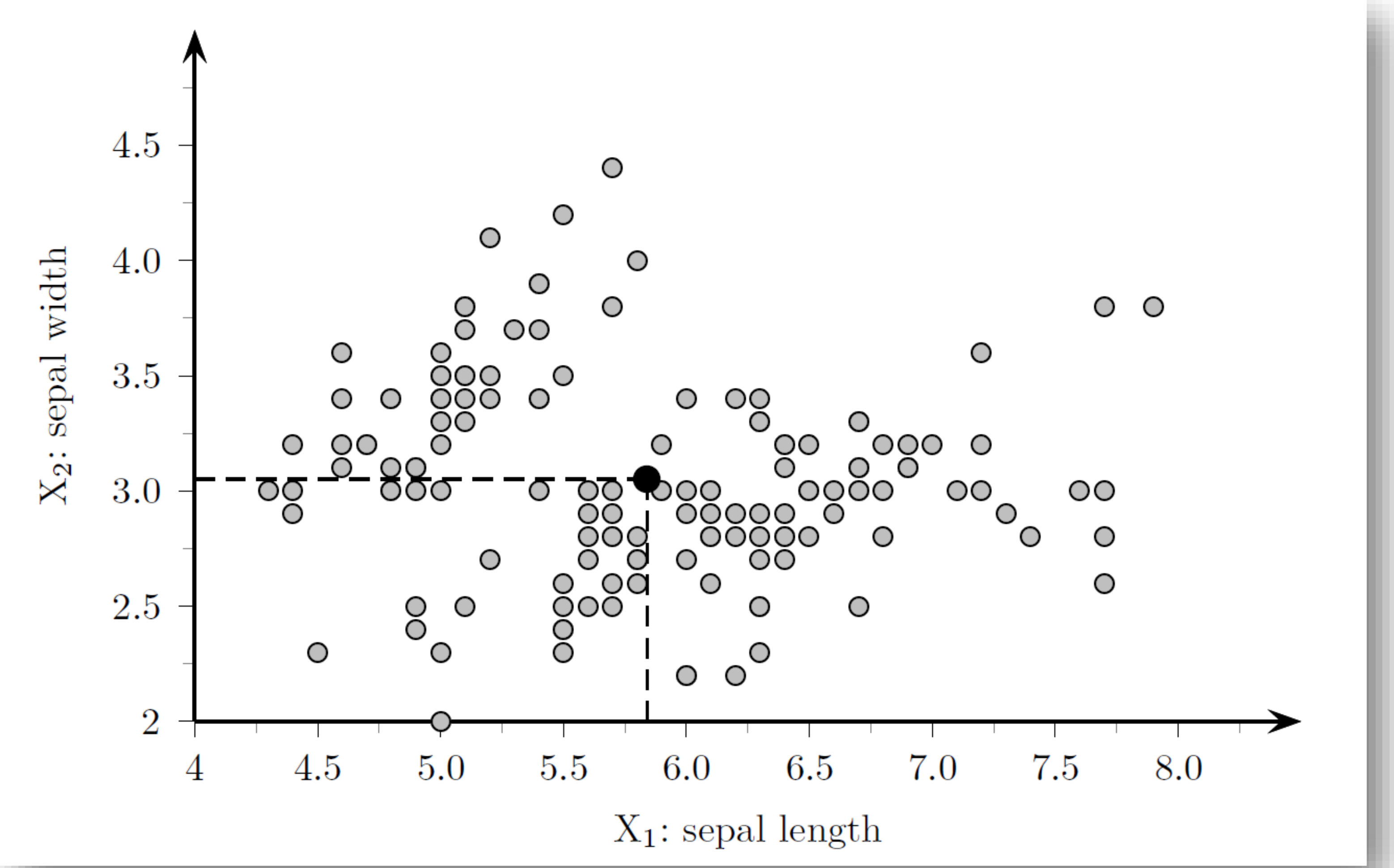
- Each entity is an Iris flower, and the attributes include sepal length, sepal width, petal length, and petal width in centimeters, and the type or class of the Iris flower.

Data Matrix

	Sepal length X_1	Sepal width X_2	Petal length X_3	Petal width X_4	Class X_5
$\mathbf{x}_1$	5.9	3.0	4.2	1.5	Iris-versicolor
$\mathbf{x}_2$	6.9	3.1	4.9	1.5	Iris-versicolor
$\mathbf{x}_3$	6.6	2.9	4.6	1.3	Iris-versicolor
$\mathbf{x}_4$	4.6	3.2	1.4	0.2	Iris-setosa
$\mathbf{x}_5$	6.0	2.2	4.0	1.0	Iris-versicolor
$\mathbf{x}_6$	4.7	3.2	1.3	0.2	Iris-setosa
$\mathbf{x}_7$	6.5	3.0	5.8	2.2	Iris-virginica
$\mathbf{x}_8$	5.8	2.7	5.1	1.9	Iris-virginica
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\mathbf{x}_{149}$	7.7	3.8	6.7	2.2	Iris-virginica
$\mathbf{x}_{150}$	5.1	3.4	1.5	0.2	Iris-setosa

https://archive.ics.uci.edu/ml/machine-learning-databases/iris/iris.data

Geometric Representation



Distance and Angle

- **Vectors**: Treating data objects and attributes as vectors, and the entire dataset as a matrix, enables to apply both geometric and algebraic methods to aid in the data mining and analysis tasks.

- What is a distance / similarity?

Dot Product

$$\mathrm{a} = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_m \end{pmatrix} \qquad \mathrm{b} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{pmatrix}$$

$$\begin{aligned} \mathrm{a}^{\mathrm{T}}\mathrm{b} &= \begin{pmatrix} a_1 & a_2 & \cdots & a_m \end{pmatrix} \times \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{pmatrix} \\ &= a_1 b_1 + a_2 b_2 + \cdots + a_m b_m \\ &= \sum_{i=1}^{m} a_i b_i \end{aligned}$$

Euclidean Norm

$$\|\mathrm{a}\| = \sqrt{\mathrm{a}^{\mathrm{T}}\mathrm{a}} = \sqrt{a_1^2 + a_2^2 + \cdots + a_m^2} = \sqrt{\sum_{i=1}^{m} a_i^2}$$

- The Euclidean norm is a special case of a general class of norms, known as L_p-*norm* (*p=2*), defined as

$$\|\mathrm{a}\|_p = \left(|a_1|^p + |a_2|^p + \cdots + |a_m|^p\right)^{\frac{1}{p}} = \left(\sum_{i=1}^{m} |a_i|^p\right)^{\frac{1}{p}}$$

Euclidean Distance

- **Euclidean distance** is the distance in Euclidean space.

$$\delta(\mathrm{a}, \mathrm{b}) = \|\mathrm{a} - \mathrm{b}\| = \sqrt{(\mathrm{a} - \mathrm{b})^{\mathrm{T}}(\mathrm{a} - \mathrm{b})} = \sqrt{\sum_{i=1}^{m}(a_i - b_i)^2}$$

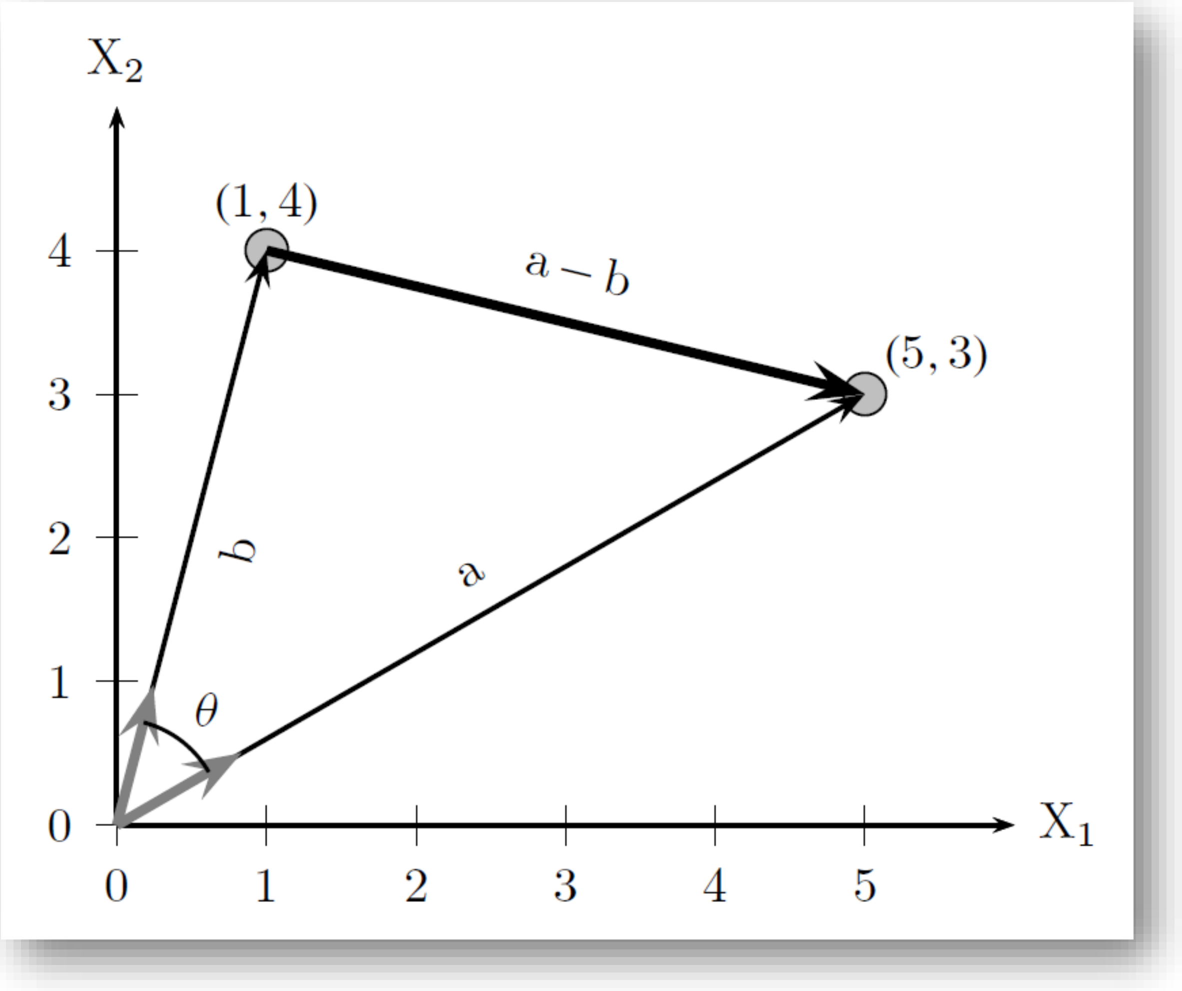
Cosine Similarity

- **Cosine similarity** is the cosine of the smallest angle between vectors *a* and *b*.

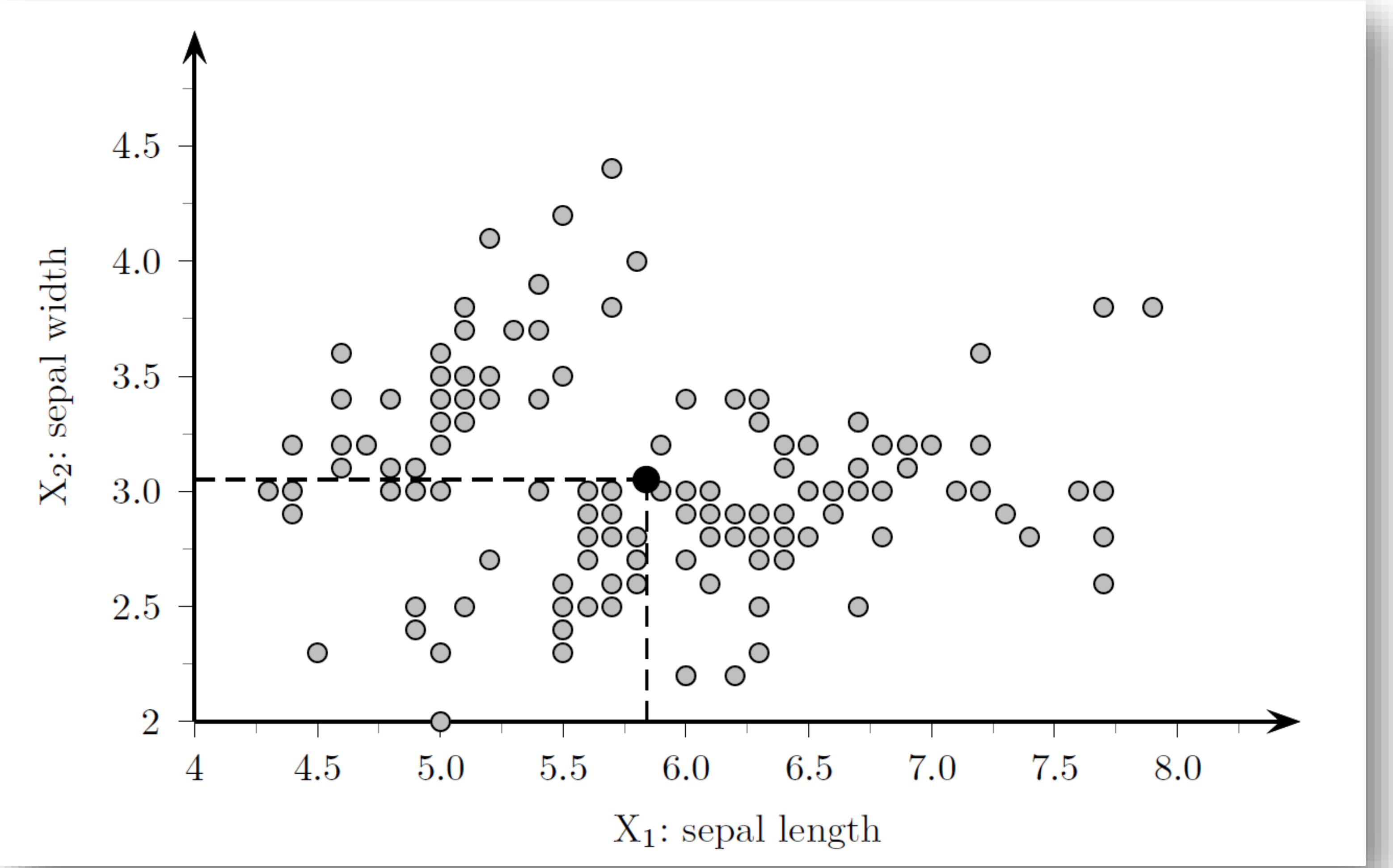
$$\cos\theta = \frac{\mathrm{a}^{\mathrm{T}}\mathrm{b}}{\|\mathrm{a}\|\,\|\mathrm{b}\|} = \left(\frac{\mathrm{a}}{\|\mathrm{a}\|}\right)^{\mathrm{T}}\left(\frac{\mathrm{b}}{\|\mathrm{b}\|}\right)$$

- The cosine of the angle between a and b is given as the dot product of the unit vectors *a / ||a||* a *b / ||b||*.

Distance and Angle



Mean and Total Variance



Mean

- The **Mean** of the data matrix D is the vector obtained as the average of all the points.

$$mean(\mathrm{D}) = \boldsymbol{\mu} = \frac{1}{n}\sum_{i=1}^{n} \mathrm{x}_i$$

Total Variance

- The **Total Variance** of the data matrix *D* is the average squared distance of each point from the mean.

$$var(\mathrm{D}) = \frac{1}{n}\sum_{i=1}^{n}\delta(\mathrm{x}_i, \boldsymbol{\mu})^2 = \frac{1}{n}\sum_{i=1}^{n}\|\mathrm{x}_i - \boldsymbol{\mu}\|^2$$

Centered Data Matrix

- Often we need to center the data matrix by making the mean coincide with the origin of the data space.

- The centered data matrix is obtained by subtracting the mean from all the points.

$$Z = D - 1 \cdot \boldsymbol{\mu}^T = \begin{pmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_n^T \end{pmatrix} - \begin{pmatrix} \boldsymbol{\mu}^T \\ \boldsymbol{\mu}^T \\ \vdots \\ \boldsymbol{\mu}^T \end{pmatrix} = \begin{pmatrix} x_1^T - \boldsymbol{\mu}^T \\ x_2^T - \boldsymbol{\mu}^T \\ \vdots \\ x_n^T - \boldsymbol{\mu}^T \end{pmatrix} = \begin{pmatrix} z_1^T \\ z_2^T \\ \vdots \\ z_n^T \end{pmatrix}$$

Linear Independence

- The vectors $v_1, . . . , v_k$ are linearly dependent if at least one vector can be written as a linear combination of the others.

- A set of vectors is linearly independent if none of them can be written as a linear combination of the other vectors in the dataset.

- **Basis**: It is a minimum set of vectors in the dataset that are linearly independent. Any two bases must have the same number of vectors.

Dimensionality

- **Dimension (rank)**: For any matrix, the *dimension* of its row and column space is the same, and this dimension is also called the *rank* of the matrix. Rank gives an indication about the intrinsic dimensionality of the data.

- **Dimensionality reduction**: It is often possible to approximate dataset (matrix) *D* with a derived data matrix D’, which has much lower dimensionality.